PROBLEM SOLVING IN STATISTICS*

by

Charles E. McCulloch
Biometrics Unit, Cornell University, Ithaca, NY 14853

BU-816-M** May 1983

Abstract

Problem-solving skills are necessary to be proficient in many areas of statistics, yet the training of these skills is absent or under-emphasized. Some general problem-solving strategies that are useful in statistics are presented and illustrated. Integration of the teaching of these skills in a statistics program is discussed.

---

* Paper presented in the Statistical Education Section of the American Statistical Association Summer Meetings, Toronto, Canada, August 1983.

** In the Biometrics Unit Series, Cornell University, Ithaca, NY 14853.

I. Introduction

Problem-solving skills are widely regarded as being under-emphasized in education today. A recent report on education in the United States (New York Times, January 9, 1983) was headlined with an article titled, "Teaching to Think: A New Emphasis. Schools and Colleges Design New Programs to Counter Indicated Drop in Reasoning and Problem-Solving Skills." The report went on to say that an emphasis on basic skills rather than applications of the skills was a possible cause. In another report (New York Times, April 24, 1983) it was stated that quantitative reasoning and numerical problem-solving skills are "... increasingly necessary to be a successful student in the liberal arts, to be successful in one's professional life and to be a successful participant in democratic society."

As important as problem solving is in society, it is perhaps even more important in statistics. It is necessary for doing research, learning to do theory, learning the application of statistical methods, in consulting and in statistical practice. Unfortunately, explicit emphasis on it in statistics programs is often limited or absent.

Programs in statistics are natural places to provide a strong emphasis on problem solving. Graduates of the field need to be equipped with the ability to approach problems in a flexible manner, whether they work in theoretical, applied or consulting capacities. Statistical service courses provide students with tools for analyzing problems and can therefore be a very natural forum for the teaching of general problem-solving skills. In this paper, we argue that the teaching of general problem-solving strategies is an effective way to increase a student's ability to solve problems.

II. Problems and problem solving

It is helpful to have some formal definitions to talk about problem solving and problem-solving strategies. Wickelgren (1974) defines a problem as consisting of three parts:

givens: Information available at the start. It can be explicitly given in the statement of the problem or implicit.

operations: Actions performed on the givens or expressions derived from the givens.

goal: Expression one wishes to end with.

A solution to a problem is defined as a listing of the operations and expressions used to get from the givens to the goals.

How do these abstract definitions fit statistical problems? Textbook problems, such as the one below from Mood, Graybill and Boes (1975) fit easily into the above framework:

> In genetic investigations one frequently samples from a binomial distribution
>
> $$f(x) = \binom{m}{x} p^x q^{m-x}$$
>
> except that observations of $x = 0$ are impossible; so, in fact, the sampling is from the conditional (truncated) distribution
>
> $$\binom{m}{x} \frac{p^x q^{m-x}}{1-q^m} I_{\{1,2,3,\cdots,m\}} .$$
>
> Find the maximum likelihood estimator in the case $m = 2$ for samples of size n. Is the estimator unbiased?

Explicit givens for this problem would be the stated distribution and the fact that $m = 2$. Implicit givens would be the meanings of maximum likelihood and unbiased. The goal is to establish the truth or falsity of the statement "The estimator is unbiased." Operations would be derived statements such as an explicit formula for the maximum likelihood estimator or its expected value.

In a problem such as the above, the goal is something to be verified or disproved. Another common type of goal is a "find" goal. This would arise in a typical methods course problem, where a data set and description form the explicit givens and the goal is to find an appropriate method and apply it. Consulting can also be formulated in this way, where the problem is to find an appropriate method or recommendation.

Though practice at solving problems is usually available in courses with homework it is usually of a fairly rigid nature, with well-defined goals and very few possible strategies to choose from for solution. In research, applications or consulting this is rarely the case and it therefore would benefit students to see flexible, general problem-solving approaches. These can be especially helpful for students that do not know where to begin a problem or are stumped after their first attempt.

III. Problem-solving needs in statistics

Where does problem solving enter into statistics? Perhaps in statistical research it is most evident that problem-solving skills are necessary. After all, research consists mainly of problem formulation and problem solving. Is any attempt made to teach students the requisite skills?

A search of the course catalog of the College of Agriculture and Life Sciences at Cornell University showed that nine of 17 departments offered research methods courses of one form or another. The Biometrics Unit does not currently offer such a course. Where then might students acquire such skills? Theory courses might conceivably offer the framework for such training but they rarely do. Often they are oriented in a theorem: proof style. This is useful for logical organization but it offers little insight into which theoretical constructs will be useful in various circumstances. Problems in theoretical

courses tend to be specific to sections of a course and offer practice only in making very limited choices. In teaching a theoretical statistics course I have found this problem with the students' math background. They have the requisite knowledge but are not able to apply it to statistics problems.

The need for problem solving in statistical methods courses has been argued for previously (Bice and Perrin, 1982; Jordan, 1982) and shown to be related to performance (Wasik, 1982). I would argue especially for problem solving in the sense of knowing what to try and how to approach difficult problems akin to Davenport's (1982) "thinking in front of a class." He also quotes Whyburn as saying "What we need is to get into the wastebaskets of the great minds and rummage through their scrap paper to see how they go about solving problems." My feeling is that a few general strategies are used over and over again and that common ones can be taught to students.

McCulloch et al. (1982) argue for problem solving in consulting and Stergion (1982) lists the desire to solve broad-base problems as a "must" for practicing industrial statisticians. In summary, problem solving is important in all areas of statistics.

Despite its importance, it is under-emphasized. Jordan (1982) proposes that statistics courses adopt "a problem-solving context that takes a broader vew of statistics than is provided in most statistics courses." Stergion (1982) states "It would help both the individual and the company involved if the universities could institute programs that involve use of statisticians in solving today's industrial problems." How can these needs be met?

IV. Some general problem-solving strategies

My experience in teaching applied and theoretical courses, in consulting and in supervising students doing consulting has led me to the following conclusions. Experienced consultants and researchers use certain general strategies over and over again to approach difficult problems almost as second nature. Students do not. Often when a direct attack on a problem fails, a student will have no ideas about what else to try.

Further, some of these strategies can be taught to students. This is not to say that all modes of approaching problems can be taught to students or even identified. However, the most common ones can.

As an example, consider the problem from Mood, Graybill and Boes (1975) listed above. I assigned this problem recently to my class as a homework problem. Many of the students were able to calculate the maximum likelihood estimator, but were not able to calculate its expected value. They then gave up.

If the students were aware of some general strategies they might then reason along the following lines: Are there other approaches that might be profitable? The student would then review possible general strategies. One of those listed below (which is often very useful) is to try to solve a simpler problem. If the student tried this technique, a logical choice would be to simplify by letting $n = 1$. For $n = 1$, the expected value is not equal to the parameter being estimated. In this example, solving a simpler problem was grandly successful. It actually solves the original problem. Usually, we only hope that solving a simpler problem will give us insight into how to solve the original problem. (A point to bring out here might be an identification of situations in which solving a simpler problem might solve the original problem.)

This sort of problem solving in mathematics is covered in great detail in the books by Polya (1957, 1962). An important point to realize is that the techniques have broader applicability. For example, a consultant faced with a problem involving a large number of variables might first solve the simpler problem with only one variable. This may give insight into how to solve the complete problem.

What are some common general strategies? Below is a list and short description of some of the techniques described in Wicklegren (1974) that I have found useful in statistics.

> Inference Draw inferences from explicitly and implicitly presented information that satisfy one or both of the following criteria:
> (a) the inferences have frequently been made in the past from the same type of information.
> (b) the inferences are concerned with properties (variables, terms, expressions and so on) that appear in the goal, the givens or inferences from the goal and givens.

This is a formal description of the usual, proceed-from-start-to-finish approach to solving a problem. It is almost always a logical first choice for a strategy.

> Classification of Action Sequences: Divide operations used to get solutions into classes. Either eliminate entire classes of operations or consider only a single class of operations at a time.

This strategy is often useful in methods courses or consulting when the problem is to find an appropriate method. The possible solutions (statistical methods) are divided into classes according to what kind of data they are appropriate for (ordinal level, two samples, etc.).

> State Evaluation and Hill Climbing: (a) choose an evaluation function and (b) pick consecutive actions so as to improve the evaluation function.

This technique is frequently used in building a regression model. The evaluation function could be the quality of the residual plot.

> Subgoals: Replace a single difficult problem with two or more simpler problems.

Hopefully, some of the subproblems are easily solvable or are analogous to previous problems. This method requires the specification of subgoals which lead to the goal. This is a good technique for problems requiring several actions to solve; not a good technique for problems requiring insight alone.

> Contradiction: Prove that a goal could not possibly be obtained from the givens when using a particular type of possible solution. Hence, a potential solution is eliminated. This method is useful in conjunction with classification of action sequences.
>
> Solving a Simpler Problem: Reduce the complexity of the current problem and solve the simpler problem. Hopefully, the insight gained in solving simpler problems allows us to solve the more complicated problem.
>
> Working Backwards: Start with the goal and try to guess a preceding statement or statements that imply the goal statement. Gradually work backwards to the givens.
>
> Solving a More General Problem: Solve the general case instead of a special case. This method tends to work if the problem can be generalized in a manner which retains its essential features, but removes unnecessary detail.

This technique is sometimes used in applied statistics. For example, rewriting a problem so that it fits a general linear model may allow one to more easily derive tests using general theory.

Some of these techniques are well illustrated by a problem I ran across while paging through an applied statistics text. Neter, Wasserman and Whitmore (1982) have a special section titled "Exact Confidence Limits for p" which gives the following interval as an exact $100(1-\alpha)\%$ confidence interval for p based on a random sample $X_1, \cdots, X_n$ from a Bernoulli distribution with parameter p $(X = \sum_{i=1}^{n} X_i)$.

$$\left[\frac{X}{X + (n-X+1)F_1} , \quad \frac{(X+1)F_2}{(X+1)F_2 + n - X} \right]$$

$$F_1 = F^{2(n-X+1)}_{2X,1-\alpha/2}$$

$$F_2 = F^{2(X+1)}_{2(n-X),1-\alpha/2} \quad .$$

At first, I refused to believe that the result could be true, since I could see no possible connection between the F and the binomial distribution. I was further frustrated by the fact that, being an applied book, it had no derivation or reference. Hence, I embarked on an attempt to verify or disprove the result.

First, I applied the technique of inference, but did not come up with much. The only fact I recalled was that exact binomial confidence intervals could be found using the statistical method, described in Mood, Graybill and Boes (1975). Looking up the result there gave the confidence interval as $[p_L, p_u]$ where p_L is the solution of

$$\alpha/2 = \sum_{x=x_o}^{n} \binom{n}{x} p_L^x (1 - p_L)^{n-x}$$

and p_u is the solution to

$$\alpha/2 = \sum_{x=0}^{x_o} \binom{n}{x} p_u^x (1 - p_u)^{n-x} \quad ,$$

where x_o = observed value of X. This did not help much.

Next, I tried working backwards. Looking at the answer, I tried to guess a preceding statement. The form of the interval endpoints was a strange configuration of F percentiles. The form of the fraction triggered the memory that the F and beta distributions were related by a similar formula. Without too much trouble I found that if $Y \sim F_n^m$ then

$$W = \frac{\frac{m}{n} Y}{1 + \frac{m}{n} Y} \sim \text{Beta}\left(\frac{m}{2}, \frac{n}{2}\right) \quad .$$

This looked promising.

Returning to the method of inference, I recalled that the beta distribution is related to the binomial distribution since it is a conjugate prior. Hence, I set two subgoals. First, relate the binomial and the beta and second, relate this to the F .

Looking back at the formulas for the statistical method led me to the conclusion that (for p_L) I wanted to relate

$$\sum_{x=x_o}^{n} \binom{n}{x} p_L^x (1 - p_L)^{n-x}$$

to the beta. I could not immediately see anything.

Next, I tried solving the simpler problem when $x_o = n$. In that case the sum reduced to

$$\binom{n}{n} p_L^n (1 - p_L)^{n-n} = p_L^n \quad .$$

This could be related to a beta as

$$p_L^n = \int_0^{p_L} n\, w^{n-1} dw \quad .$$

For $x_o = n - 1$ we need to relate

$$n\, p_L^{n-1}(1-p_L) + p_L^n \quad \text{to a beta} \quad .$$

Integration by parts of

$$\int_0^{p_L} n(n-1) w^{n-2} (1 - w) dw$$

gives the right answer and in general

$$\frac{\alpha}{2} = \sum_{x=x_o}^{n} \binom{n}{x} p_L^x (1 - p_L)^{n-x} = \int_0^{p_L} \frac{w^{x_o - 1}(1-w)^{n-x_o}}{B(x_o, n-x_o+1)}\, dw \quad ,$$

i.e., p_L is the $\alpha/2$-percentile from a Beta$(x_0, n - x_0 + 1)$. p_u can be handled in a similar fashion. This completes the first subgoal.

Next, I related the F and beta percentiles. If $W \sim \text{Beta}(a,b)$, then

$$P\{W \le w_\gamma\} = \gamma$$

is equivalent to

$$P\left\{\frac{\frac{2a}{2b} Y}{1 + \frac{2a}{2b} Y} \le w_\gamma\right\} = \gamma$$

where $Y \sim F^{2a}_{2b}$. Rearranging the above gives that

$$w_\gamma = \frac{\frac{2a}{2b} F^{2a}_{2b,\gamma}}{1 + \frac{2a}{2b} F^{2a}_{2b,\gamma}} \quad .$$

Plugging in $a = x_0$ and $b = n - x_0 + 1$ finishes things off and shows that

$$p_L = \frac{\frac{2x_0}{2(n-x_0+1)} F^{2x_0}_{2(n-x_0+1),\alpha/2}}{1 + \frac{2x_0}{n-x_0+1} F^{2x_0}_{2(n-x_0+1),\alpha/2}}$$

$$= \frac{x_0}{(n-x_0+1)F^{2(n-x_0+1)}_{2x_0,1-\alpha/2} + x_0} \quad .$$

The solution for p_u proceeds similarly.

As a side remark, this technique does not give exact intervals in the sense that

$$P_p\{p_L \le p \le p_u\} \equiv 1 - \alpha \quad ;$$

however, it does give intervals with the property that

$$P_p\{p_L \le p \le p_u\} \ge 1 - \alpha$$

and usually (most n, α)

$$\inf_p P_p\{p_L \le p \le p_u\} = 1 - \alpha \quad .$$

V. Integration of problem solving

General problem-solving strategies can be incorporated easily into course-work in statistics programs. It is mainly a matter of emphasis. As Davenport (1982) recommends, thinking out loud in front of the class can be very useful. Here some mention of the general strategy being used is useful. In theory courses it may not be best to present the slickest proof possible. Though logically tidy, it does not lend insight as to how to approach a theoretical problem or research. It may also be helpful to remember that one way of learning when a method is useful is to learn when it does not work.

General strategies can usefully be gone over when reviewing homework solutions. Again, a change of emphasis is needed away from the actual solution towards why that method of solution was chosen. For more emphasis, tutorials could be given in a discussion section.

In statistical methods courses, the strategies can be emphasized when doing case studies, again by thinking out loud. For research, problem-solving strategies can be discussed and tried out in an informal seminar or journal-club atmosphere. None of these recommendations requires substantial curriculum changes.

VI. Summary

This paper proposes a slight shift of emphasis in statistics coursework away from concepts and techniques towards flexible strategies for applying the concepts and techniques to problems. A list of possible strategies for students to try when stumped on a problem are described and illustrated. Ways of integrating these techniques into the statistics curriculum are discussed.

References

Bice, T. W. and Perrin, E. B. (1982). "Teaching of Statistics in Health Services Administration," Proceedings of the Statistical Education Section of the ASA, 7-8.

Davenport, J. M. (1982). "Some Simple Things that Aid Teaching of the Art of Statistics," Proceedings of the Statistical Education Section of the ASA, 43-47.

Jordan, E. W. (1982). "A Systems Perspective of Statistics," Proceedings of the Statistical Education Section of the ASA, 37-41.

McCulloch, C. E., Boroto, D. R., Meeter, D., Polland, R. P. and Zahn, D. A. (1982). "A Holistic Approach to Training Statistical Consultants," Proceedings of the Statistical Education Section of the ASA, 116-121.

New York Times, January 9, 1983, "Teaching to Think: A New Emphasis. Schools and Colleges Design New Programs to Counter Indicated Drop in Reasoning and Problem-Solving Skills."

New York Times, April 24, 1983, "New Priority: Technological Literacy."

Polya, G. (1957). How to Solve It (2nd Ed.), New York: Doubleday.

Polya, G. (1962). Mathematical Discovery, Vol. 1: On Understanding, Learning and Teaching Problem Solving, New York: Wiley.

Stergion, A. P. (1982). "Industry's Expectations of B.S. Statisticians," Proceedings of the Statistical Education Section of the ASA, 67-69.

Wasik, J. L. (1982). "The Relationship between Problem-Solving Ability and Statistics Achievement," Proceedings of the Statistical Education Section of the ASA, 159-161.

Wickelgren, W. A. (1974). How to Solve Problems, San Francisco: Freeman.